

Analysis of the Squat Exercise from Visual Data

Fatma Youssef¹ ^a, Ahmed B. Zaky^{2,3} ^b and Walid Gomaa^{4,5} ^c

¹Computer Science Engineering, Egypt-Japan University of Science and Technology, Alexandria, Egypt

²Computer Science and Information Technology Programs (CSIT),
Egypt Japan University of Science and Technology, Egypt

³Shoubra Faculty of Engineering, Benha University, Benha, Egypt

⁴Cyber Physical Systems Lab, Egypt Japan University of Science and Technology, Egypt

⁵Faculty of Engineering, Alexandria University, Alexandria, Egypt

Keywords: Exercise Assessment, Deep Learning, Transfer Learning.

Abstract: Squats are one of the most frequent at-home fitness activities. If the squat is performed improperly for a long time, it might result in serious injuries. This study presents a multiclass, multi-label dataset for squat workout evaluation. The dataset collects the most typical faults that novices make when practicing squats without supervision. As a first step toward universal virtual coaching for indoor exercises, the main objective is to contribute to the creation of a virtual coach for the squat exercise. A 3d position estimation is used to extract critical points from a squatting subject, then placed them in a distance matrix as the input to a multi-layer convolution neural network with residual blocks. The proposed approach uses the exact match ratio performance metric and is able to achieve 94% accuracy. The performance of transfer learning as a known machine learning technique is evaluated for the squat activity classification task. Transfer learning is essential when changing the setup and configuration of the data collection process to reduce the computational efforts and resources.

1 INTRODUCTION

Physical activity is critical for our overall health. Regular exercise strengthens our muscles and bones, making daily tasks simpler. It reduces the risk of heart disease and aids in the maintenance of normal blood pressure. The impact of regular exercise on our immune system is also noticeable. Exercise also aids in the reduction of stress, anxiety, and depression (Abou Elmagd, 2016; Reiner, Niermann, Jekauc, & Woll, 2013; Garber et al., 2011; Ohuruogu, 2016; Schuch et al., 2016).


Due to the COVID-19 pandemic, it is now normal to undertake workouts at home (Kaur, Singh, Arya, & Mittal, 2020). Squat is one of the workouts that have a significant impact on muscular building and can be done at home because it requires no equipment or a lot of space and is just dependent on one's weight (Füzéki, Groneberg, & Banzer, 2020).


However, exercising at home without supervision


may be risky and lead to serious injury, especially for strength exercises such as squats. Performing squats incorrectly several times raises the risks of lower limbs and trunk injury (Lorenzetti et al., 2018). Thus, it is necessary to have an automated software system to monitor people while working out at home. The model should provide users with immediate feedback on their performance and mistakes. This system's goal is to assist individuals in preventing injuries when performing squats and to ensure that they get the most out of their fitness workout (Liao, Vakaniski, & Xian, 2020; Ogata, Simo-Serra, Iizuka, & Ishikawa, 2019).

There are various techniques to monitor individuals while they undertake workouts using sensing modalities such as RGB cameras, inertial measurement sensors, depth cameras, infrared cameras, etc. However, the most accessible and affordable way to monitor people while performing exercises is by using a single monocular camera that can be found on all mobile phones (Strömbäck, Huang, & Radu, 2020; Ogata et al., 2019).

In this work, the main objective is to assess the

^a  <https://orcid.org/0000-0002-6286-9698>

^b  <https://orcid.org/0000-0002-3107-5043>

^c  <https://orcid.org/0000-0002-8518-8908>

quality of squat exercises and identify the type of mistake(s), if any, in the performed squat. Despite the numerous datasets in the literature related to activity recognition in general, such as UCF101 Dataset (Soomro, Zamir, & Shah, 2012) and HMDB Dataset (Kuehne, Jhuang, Garrote, Poggio, & Serre, 2011), or particularly to exercise recognition, such as the MM-Fit Dataset (Strömbäck et al., 2020), there are a few datasets for exercise assessment such as KIMORE Dataset (Capecci et al., 2019), Functional Senior Fitness dataset (Bernardino et al., 2016) and Single Individual Dataset (Ogata et al., 2019). KIMORE Dataset (Capecci et al., 2019) and Functional Senior Fitness dataset (Bernardino et al., 2016) aim to assess the quality of rehabilitation exercises and the fitness level of elderly people respectively. While Single Individual Dataset (Ogata et al., 2019) assesses the squat exercise in particular into seven categories, six of them for bad squat and one for good squat. The dataset contains only one subject performing squats to produce a specific type of mistake each time.

The lack of datasets for squat assessment motivated us to develop a dataset called EJUST-SQUAT-21. EJUST-SQUAT-21 Dataset is an extension of the work presented by Fayez, Sharshar, Hesham, Eldifrawi, and Gomaa (2022) by increasing the number of video samples and adding proper annotation for various classification tasks. The details of the dataset will be described in section 3.

In contrast to the previous datasets, the EJUST-SQUAT-21 Dataset features a variety of users practicing squats without being instructed on how to do them correctly; consequently, the errors in their squats are natural. Since users do squats without instruction, it is usual for them to make many mistakes, producing the EJUST-SQUAT-21 Dataset multiclass multilabel.

The videos in EJUST-SQUAT-21 Dataset had a substantial difference in clothing and lighting circumstances, so the proposed approach uses pose estimations to minimize these dependencies and create a skeleton pose for the user executing squats. The Ogata et al. (2019) concept is used to encode each video's skeletal positions in a distance matrix that represents the Euclidean distance between key points. In order to classify the mistake in the performed squat, the computed distance matrix of each video is used as an input to a convolution neural network described in section 4.2.

Various publicly accessible benchmark datasets are used in addition to our dataset. Transfer learning is applied across various datasets to evaluate our system's robustness to various setups and configurations while also decreasing the computing resources needed for end-to-end training. Our technique can run

in real-time, so it may be embedded in mobile devices and provide immediate feedback to users while they execute squat workouts. The following are our main contributions:

- Expanding the work of Fayez et al. (2022) by concentrating on the visual data acquired from participants performing squats with the goal of increasing the amount of video samples and creating suitable annotations in order to use the dataset for different levels of squat evaluation.
- Using the collected dataset to accurately classify the squat workout error.
- Operating and validating transfer learning for different squat classification models across several datasets. Review activity recognition and squat workout evaluation.

The paper is organized as follows: Section 1 is an introduction. Section 2 describes the related work in pose estimation, activity recognition, and exercise assessment. Section 3 describes the collected dataset. Section 4 introduces the proposed approach and deep neural network architecture. The experimental work done to measure the model performance on the EJUST-SQUAT-21 Dataset and the utilization of transfer learning using different datasets is detailed in section 5. Section 6 summarizes the achieved results of the performed experiments. Finally, Section 7 concludes the paper by referring to future research directions that may be pursued.

2 RELATED WORK

2.1 Pose Estimation

Many tracks of research in activity recognition and exercise evaluation rely on the skeleton pose of the person as input to the model to avoid any dependencies on human clothes, lighting conditions, or background. Hence, the effectiveness of classification depends on the extracted **skeletal pose**. As a result, it is critical to utilize a very precise and reliable pose estimate approach.

There are many 2d and 3d pose estimations in the literature. OpenPose (Cao, Hidalgo, Simon, Wei, & Sheikh, 2019) is a powerful 2d pose estimation method. The 2d keypoints locations extracted by OpenPose can be utilized to generate 3d pose estimates (Martinez, Hossain, Romero, & Little, 2017). In addition, the 3d Pose estimates can be generated using only monocular RGB camera (Mehta et al., 2017; Popa, Zanfir, & Sminchisescu, 2017). Moreover, the shape and 3d pose estimates can be directly

extracted from raw images (Kanazawa, Black, Jacobs, & Malik, 2018a). Bazarevsky et al. (2020) proposed BlazePose 3d pose estimation which generate 33 keypoints in **real-time**. The BlazePose model consists of an encoder-decoder network for heatmaps and an encoder network to regress the keypoints coordinates. The heatmap branch is only used during training which makes the model inference time small. The model is trained on a large collected dataset for fitness exercises. It outperforms OpenPose (Cao et al., 2019) pose estimation in terms of accuracy and speed.

2.2 Activity Recognition based on Pose Estimation

The action recognition task is inextricably linked to pose estimation. Many researchers used posture estimation to improve activity identification accuracy and vice versa. Using a spatial-temporal And-Or graph model, Nie, Xiong, and Zhu (2015) developed an approach that integrates the classification of the performed action with the estimate of body parts' locations while doing that activity. Their method outperformed previous works, confirming the strong relation between pose estimation and action recognition. Iqbal, Garbade, and Gall (2017) proposed a pictorial structure model for pose estimation and action classification. By starting with uniform action prior, they used the pose estimated in each frame to update the prior, then utilized the updated prior for generating refined poses.

Posture-based Convolution Neural Network descriptors (P-CNN) were suggested by Chéron, Laptev, and Schmid (2015) for action recognition, and they used pose estimation. Patches of human body parts, as well as their optical flow, are sent into P-CNN. RGB CNN and flow CNN, respectively, are supplied with them. After aggregation and normalization, the output of each branch is coupled to construct a P-CNN descriptor that is trained using linear SVM to categorize the action done.

Strömbäck et al. (2020) introduced MM-Fit Dataset, which consists of a collection of data from different modalities such as inertial sensor data collected from smartwatches, smartphones, and earbuds, 2d and 3d pose estimates from a Multiview RGB-D camera. They trained each modality separately using a stacked convolution autoencoder to extract the important features from each modality. For inertial data, the autoencoder architecture consists of 1D convolution. For the collected videos, they used 3d pose estimates from Martinez et al. (2017) and used cylindrical coordinates as suggested by Ke, Bennamoun, An, Sohel, and Boussaid (2017) to be input to the

autoencoder because it is less sensitive to the background and illumination variation. They arranged the data of 3d pose estimation into 3d images. The autoencoder architecture used for the 3d pose estimates employed 2d convolutions instead of 1d convolutions. After training each unimodal, they stacked the latent layer representation of each to train a fully connected multimodal autoencoder. Then, the latent layer of this latter autoencoder is used as input for fully connected network to classify between 10 different exercises, which are squats, lunges, bicep curls, sit ups, pushups, and tricep extensions, dumbbell rows, jumping jacks, dumbbell shoulder press, and lateral shoulder raises. The whole architecture is trained end-to-end to perform this classification. The proposed model utilizes the 3d pose estimates convolution autoencoder branch in transfer learning between the MM-Fit Dataset and EJUST-SQUAT-21 Dataset, as described in Section 5.2.

2.3 Exercise Assessment

The main goal of our research is to classify squat exercise performance. One method is to use Kinect to retrieve the key points of the subject's skeleton stance while executing squats, then compute the joint angles and compare them to the standard using the program proposed by Vybornyi, Rozaliev, and Orlova (2017). This issue is closely related to the evaluation of physical rehabilitation. Liao et al. (2020) presented a deep neural network model that takes as input a series of skeletal joint positions recorded by the sensory system for a person executing 10 workouts such as deep squats. Sub-convolutional networks are used to handle spatial information in joint displacement, while recurrent layers are used to capture temporal relationships. They used Gaussian Mixture Model (GMM) log-likelihood as a performance metric. The network outputs a prediction for the quality score of the performed exercise.

To distinguish between squatting, standing, and stopping, Hung, Liu, and Chang (2020) employed a convolution neural network. Depth cameras and a marker-based motion tracking system are used to capture the input data. The latter was utilized as a benchmark for determining classification accuracy. Ogata et al. (2019) collected a dataset using single ordinary camera; then keypoints were extracted using 3d pose estimation (Kanazawa, Black, Jacobs, & Malik, 2018b). The extracted keypoints are used to compute the distance matrix for each video where each column in the matrix captures the distance between keypoints in each frame. The distance matrix is considered as the input to the model architecture, which is

composed of multiple convolution layers and residual blocks (He, Zhang, Ren, & Sun, 2016). The output of the network is a softmax layer with 7 neurons corresponding to 7 different classes of squat error. The proposed model uses the same architecture with some modifications described in Section 4.2 to train our model.

3 DATASETS

3.1 EJUST-SQUAT-21 Dataset

The EJUST-SQUAT-21 Dataset is an expansion of the dataset collected in (Fayez et al., 2022). Their gathered dataset is a bimodal dataset of 27 participants executing squats utilizing both visual and inertial data streaming. The majority of the respondents in the sample are men, except the exception of 3 African ladies, who are recorded while exercising for cultural reasons. The respondents' ages range from 18 to 25 years old. Except for the female recordings, which were taken indoors at the gymnasium hall, most of the data were obtained outdoor at the Egypt Japan University of Science and Technology Campus. The subjects were asked about their level in performing squats, and their answers varied from beginners to experts. Each subject in the dataset was asked to wear 9 IMU sensors distributed among his body, and perform squats around 6 times in two recording sessions without any supervision to make the dataset representable of the common mistakes that the subject may make at home without the presence of a trainer. The difference between the recording sessions of the subject is only in his direction with respect to the camera, either facing the camera (frontal view) or side (lateral view). The camera used in the data collection is RGB mobile camera.

The main objectives of the EJUST-SQUAT-21 Dataset are utilizing the visual data gathered (Fayez et al., 2022) to increase the number of videos available for training a classification model and creating relevant annotations for the videos so that they could be used to assess the performed squat on various levels. In order to fulfill the first objective, each video is divided into many shorter ones, each including only one squat repetition. This necessitates knowledge of the start frame of each squat repetition, as well as the original video's cutting points. The subject is in a standing posture at the start of each squat repetition. As a result, his knee angles exceed 150 degrees. Calculating the subject's knee angle in each frame and retrieving the frame numbers where the subject is in a standing position, the model uses the extracted joint

locations of BlazePose (Bazarevsky et al., 2020) pose estimation to calculate the knee angle. By representing each squat exercise in one video, the number of videos in the dataset is increased, and their length is made equal, 50 frames per video, to be suitable as input to our model described in section 4. Figure 1 shows examples of EJUST-SQUAT-21 Dataset with its annotation.

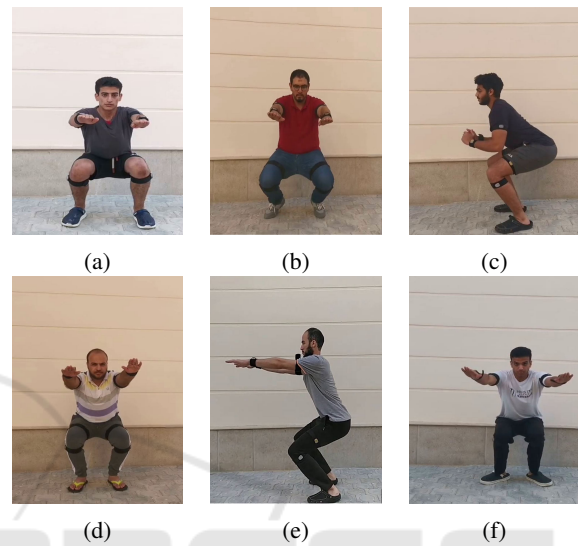


Figure 1: (a) Good Squat. (b) Bad Squat (Heels move off the ground). (c) Bad Squat (Hips are not parallel to the ground). (d) Bad Squat (Distance between feet is less than the distance between shoulders). (e) Bad Squat (Knees exceed feet and heels move off the ground). (f) Bad Squat (Foot externally rotated and Inward knees).

Annotation of each video in the dataset is created with the help of a professional trainer to clarify the type of mistake(s), if any, in the performed squat. Six common mistakes are discovered: less space between the feet than shoulders, knees beyond the feet, heels off the ground, incorrect foot rotation, hips not aligned with the ground, or inward knees.

In order to achieve our second objective, three distinct annotations are given for three different levels of squat evaluations. First, EJUST-SQUAT-21 Dataset videos are divided into two 2 classes: excellent squats (no mistakes) and poor squats (at least one of the mistakes listed above). With this annotation, we are able to test our model's capacity to discriminate between good and bad squats. Second, the videos are divided into three 3 classes based on the degree of the subject performing squats: beginner (doing squats with several mistakes), intermediate (doing squats with one minor mistake), and expert (squatting without any mistakes). Third, the dataset is labeled to specify the type of mistake made. The dataset has 7 different labels, 6 for a bad squat and one for a good squat. These

different labels are shown in Table 1 with the number of videos in each label. There may be more than one mistake in the squat exercise. Therefore, EJUST-SQUAT-21 is a multi-label dataset. Table 2 shows the number of videos with their corresponding multi-label.

Table 1: Description of Labels in EJUST-SQUAT-21 Dataset.

Label	Label Description	Number of videos
1	Distance between feet is less than distance between shoulders	87
2	Knees exceed feet	98
3	Heels move off ground	78
4	Foot externally rotated	15
5	Hips are not parallel to the ground	139
6	Inward knees	54
7	Good	200

Table 2: Labels in EJUST-SQUAT-21 Dataset.

Label 1	Label 2	Label 3	Label 4	Label 5	Label 6	Label 7	Total number of Videos
0	0	0	0	0	0	1	200
0	0	0	0	0	1	0	32
0	0	0	0	1	0	0	100
0	0	0	0	1	1	0	17
0	0	0	1	0	0	0	10
0	0	0	1	0	1	0	5
0	1	0	0	0	0	0	8
0	1	1	0	0	0	0	56
0	1	1	0	1	0	0	22
1	0	0	0	0	0	0	75
1	1	0	0	0	0	0	12

3.2 More Datasets for More Efficiency

The effectiveness of the proposed approach is proven by using three datasets. EJUST-SQUAT-21 Dataset, Single Individual Dataset (Ogata et al., 2019), and MM-Fit Dataset (Strömbäck et al., 2020), to measure the effectiveness of transfer learning between them.

The Single Individual Dataset consists of a single individual performing squat to order to make a specific error. There are seven distinct classes in the dataset, six of which are connected to bad squat and one to good squat. Inward knees, round backs, wrapped backs, upwards heads, shallowness, and frontal knees are the six forms of squat errors.

The MM-Fit Dataset is a multimodal dataset for single-person activity recognition. The modalities used in the dataset are 2d and 3d pose estimates from RGB-D camera and accelerator and gyroscope from smartwatches, earbuds, and smartphones. The subjects in the dataset performed 10 different exercises: squats, sit-ups, lunges, sitting dumbbell shoulder press, bicep curls, push-ups, jumping jacks, standing dumbbell rows, dumbbell lateral shoulder raises, and sitting overhead dumbbell triceps extensions.

4 PROPOSED APPROACH

The proposed method is based on the work presented by Ogata et al. (2019). Figure 2 shows an overview

of the whole architecture. Given the video of the subject performing squats, the skeletal pose of the subject is extracted for each frame using BlazePose pose estimation (Bazarevsky et al., 2020). Using the keypoints in the skeletal pose, a difference matrix representing the Euclidean distance between each pair of joints in each frame is computed. The difference matrix used as an input to a convolution neural network that performs a binary, trinary, or the seven classes. The aim of classification is distinguishing between good and bad squats, determining the level of the subject performing squats, whether it is beginner, intermediate or advanced, or specifying the mistake(s) in the performed squats.

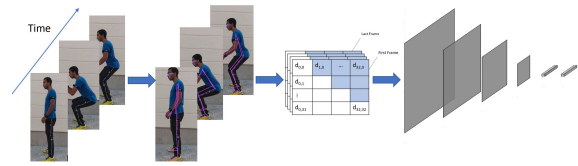


Figure 2: Overview of the whole architecture.

4.1 3D Pose Estimation and Difference Matrix

The pose estimation technique described in BlazePose (Bazarevsky et al., 2020) is used. It generates 33 keypoints for a single person in real-time. Figure 3 shows the generated keypoints.

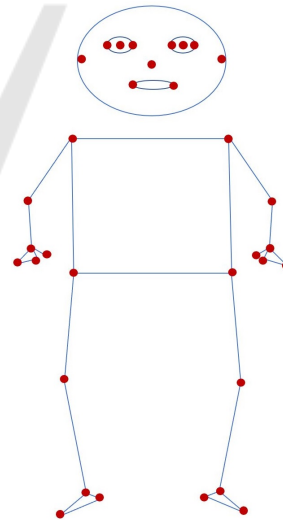


Figure 3: Keypoints generated by BlazePose (Bazarevsky et al., 2020).

For each frame, BlazePose pose estimation generates 33 keypoints, they are arranged into a difference matrix where each element $d_{i,j}$ equals the Euclidean

distance between joint i and joint j .

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

So, each frame will be represented in a 33×33 matrix. The unique elements in each matrix can be found in its upper triangular or lower triangular, so there are only $(33 * 32)/2 = 528$ unique elements in each matrix. By flattening (vectorizing) the unique elements of each matrix in a column vector and stacking them together, the final distance matrix for a video will have dimensions of $528 \times N$, where N is the number of frames in a video. Figure 4 shows the construction of the difference matrix where the shaded elements are the unique elements in each frame.

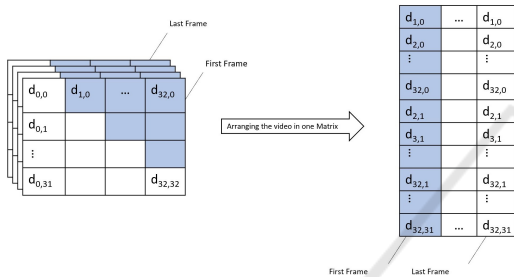


Figure 4: Difference Matrix.

4.2 Network Architecture

The same architecture shown in (Ogata et al., 2019) is used with two modifications. First, only two residual blocks instead of four blocks to reduce overfitting. Second, one additional fully connected layer followed by batch normalization and LeakyReLU activation before the output layer. The dimensions of the fully connected and output layers depend on the classification task of the performed squat. In the case of training on EJUST-SQUAT-21 Dataset, two neurons are used in these layers to classify the performed squats into good or bad. Three neurons are used to classify the performed squats into beginner, intermediate, and expert. Seven neurons are used to determine the mistakes in the performed squats. For any other dataset, Seven neurons are used in these layers.

5 EXPERIMENTS

The network architecture shown in Figure 5 is used and AdaDelta optimizer as in (Ogata et al., 2019) with a batch size of 32. Various experiments are performed utilizing multiple datasets.

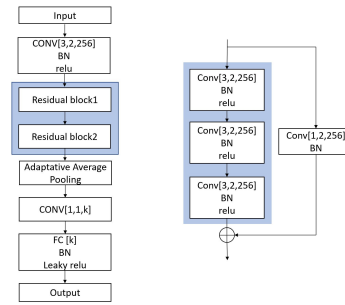


Figure 5: Network architecture where k represents the number of classes which can be 2, 3, or 7 based on the goal of classification.

5.1 EJUST-SQUAT-21 Dataset/Single Individual Dataset

To measure the performance on EJUST-SQUAT-21, four experiments are implemented. The first three experiments perform end-to-end training on EJUST-SQUAT-21 Dataset to perform binary, trinary, and multi-label classification, respectively. The goal of the fourth experiment is to measure the effect of transfer learning from the Single Individual Dataset to EJUST-SQUAT-21 Dataset. A pretrained model on the Single Individual Dataset (Ogata et al., 2019), then used the final weights as initialization for training the model on EJUST-SQUAT-21 Dataset to perform multi-label classification.

Two experiments are performed to measure the performance on the Single Individual Dataset. In the first experiment, end-to-end training is implemented. In the second experiment, a pretrained model based on EJUST-SQUAT-21 Dataset is used. As an initialization, the final weights are utilized for training/fine-tuning the model on the Single Individual Dataset to measure the effect of transfer learning from EJUST-SQUAT-21 Dataset to the Single Individual Dataset.

5.2 MM-fit/EJUST-SQUAT-21/Single Individual Datasets

Another experiment is performed utilizing the MM-Fit Dataset. As mentioned in section 2.2, the MM-Fit Dataset is a multimodal dataset where each modality is trained separately using an autoencoder, then stacked together and trained end-to-end for activity recognition. The model is implemented based on the training of the 3d pose estimation modality using autoencoder. The architecture of the autoencoder used in (Strömbäck et al., 2020) is shown in Figure 6a. the autoencoder is trained on the 3d pose estimation modality in the MM-Fit Dataset for 25 epochs to extract the features that distinguish between differ-

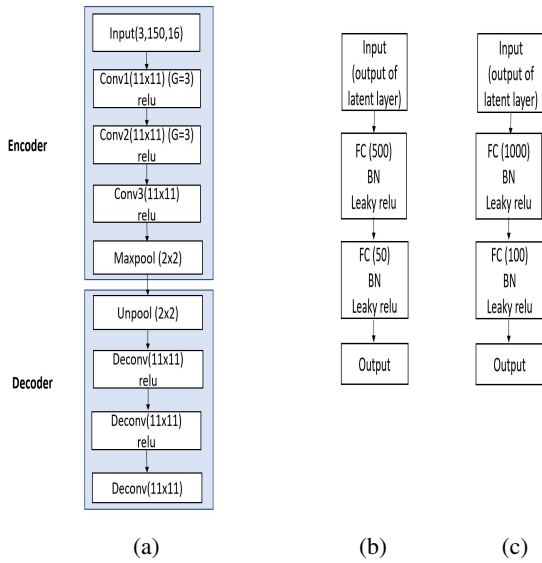


Figure 6: (a) MM-Fit Architecture. (b) EJUST-SQUAT-21 Dataset Classifier. (c) Single Individual Dataset Classifier.

ent exercises. Then, the encoder part of the network is used to extract the features in EJUST-SQUAT-21 Dataset, as well as the Single Individual Dataset separately without training. For each dataset, the output of the latent layer is passed to a fully connected neural network to classify the mistake in the squat exercise into seven different categories. In the case of EJUST-SQUAT-21 Dataset, there may be more than one mistake in the exercise, hence, multi-label. Whereas in the Single Individual Dataset (Ogata et al., 2019) there is only one mistake in each video or no mistake (good squat); hence, single label. The classification performance of the EJUST-SQUAT-21 Dataset using exact match ratio, while using the accuracy for the Single Individual Dataset.

6 RESULTS

The classification of squats in EJUST-SQUAT-21 Dataset (two classes): good or bad, The model training is described in section 4.2 with 2 neurons in the last two layers for 100 epochs (manually fixed). In order to overcome the problem of imbalanced data, a weighted cost function is used to penalize the error in classifying bad squats twice as much as the error in classifying good squats. The model achieved 96% accuracy using 10-fold cross-validation. Figure 7 shows the resulted confusion matrix from the addition of the confusion matrix of each fold, then normalizing the final matrix.

The confusion matrix shows that most misclassification is due to misclassifying the bad squat to

good squat. That is because the number of good squat videos in the dataset is larger than the number of bad squat videos. Also, bad squat videos vary much depending on the type and number of mistakes, so the model can easily classify the bad squat with one mistake as a good one.

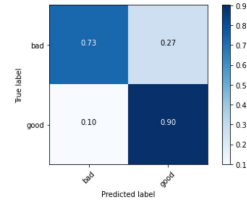


Figure 7: Normalized confusion matrix for squat classification into good or bad squat.

To classify the player's level into beginner, intermediate, and expert, the experiment mentioned above is repeated on EJUST-SQUAT-21 Dataset with 3 softmax neurons in the last layer achieving 92% accuracy. Figure 8 shows the normalized confusion matrix. The confusion matrix shows that the model is more confident in predicting the advanced class than the intermediate and beginner class because both classes contain significant variations of the type and number of errors.

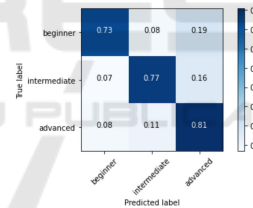


Figure 8: Normalized confusion matrix for squat player's level classification into beginner, intermediate, and expert.

Training the network architecture shown in Figure 5 using EJUST-SQUAT-21 Dataset for 83 epochs with early stopping criterion to classify the type of mistake(s) in squat, if any, the model achieved 94% accuracy using 5-fold stratified cross-validation. The stratified cross-validation was used to take into account the imbalanced dataset. The ratio of exact match is used to measure the model's accuracy, considering the prediction is correct only if all the labels match the actual target. Hamming loss is used to assess the performance of our model. The average loss is 0.015.

For the rest of the paper, EJUST-SQUAT-21 Dataset is used with multi-label annotation to classify the type of the mistake in squats unless otherwise mentioned. In addition, the number of epochs used for training is determined based on early stopping criterion except otherwise mentioning.

Pretraining the model on the Single Individual Dataset, then fine-tuning on EJUST-SQUAT-21 yields an accuracy of 90% in only 38 epochs compared to an accuracy of 94% in case of end-to-end training for 83 epochs.

Whereas pretraining on the MM-Fit Dataset achieves 88% accuracy in 200 epochs (manually fixed) using the following approach. First, the autoencoder architecture shown in Figure 6a is used to pretrain the 3d pose estimation modality in MM-Fit Dataset (Strömbäck et al., 2020). Second, the encoder part of the network is used as a feature extractor for EJUST-SQUAT-21. Third, the latent layer output is passed to the classification network shown in Figure 6b. Table 3 summarizes the results achieved on EJUST-SQUAT-21 Dataset.

In comparison to end-to-end training, pretraining on the Single Individual Dataset has a significant impact, as it allows the EJUST-SQUAT-21 Dataset to reach its best accuracy in less than half the number of epochs with a loss of just 4% of accuracy. The high similarity between these datasets is the reason behind this. They are both for squat evaluation and classification into 6 types of mistakes or no mistakes, with some common errors in the datasets like inward knees and frontal knees.

Compared to the significant effect of pretraining on Single individual Dataset, pretraining the model on MM-Fit Dataset has a poor effect on the model's accuracy because MM-Fit Dataset is specified for exercise recognition in general, not particularly squat.

The task of classifying Single Individual Dataset into 7 disjoint classes that represent the type of mistake, end-to-end training is implemented using the model architecture shown in Figure 5. The dataset is divided into 60% training, 10% validation, and 30% test, The model achieved 80% accuracy in 60 epochs. Whereas the accuracy achieved when dividing the dataset into 80% training, 10% validation, and 10% test is 87%. The results show that the model needs more training data to reach higher accuracy.

The Single Individual Dataset is used for the rest of the experiments, with 80% – 10% – 10% split. When a pretraining process is applied to the model shown in Figure 5 on EJUST-SQUAT-21, then fine-tuning on the Single Individual Dataset (Ogata et al., 2019), the model achieved an accuracy of 80.2% in 54 epochs compared to 87% accuracy in case of end-to-end training for 60 epochs.

The results show that transfer learning from Single Individual Dataset to EJUST-SQUAT-21 Dataset has less impact on accuracy compared to transfer learning from EJUST-SQUAT-21 Dataset to Single Individual Dataset. The reason for that is that the classification

task for EJUST-SQUAT-21 Dataset is more complex since it is multi-label. Hence, pretraining the model on it increases its ability to perform multi-class classification on the Single Individual Dataset.

The model achieved 73% accuracy when training a simple fully connected neural network shown in Figure6a on the extracted features from the Single Individual Dataset from the encoder part of the autoencoder architecture shown in Figure 6c after being trained on the MM-Fit Dataset (Strömbäck et al., 2020).

The above results are summarized in Table 3.

7 CONCLUSIONS and FUTURE WORK

In this paper, utilization of the visual part of the dataset collected in (Fayez et al., 2022) is proposed, in addition to increasing the data samples and providing proper annotation for various levels of squat assessment. The results introduced show the effectiveness of using 3d pose estimation as input to the model. The deep model architecture described in (Ogata et al., 2019) is used with major modifications to increase the model generalizability. The approach is tested on two squat assessment datasets: EJUST-SQUAT-21, and the Single Individual Dataset. The results show that the model can achieve remarkable accuracy, even with stringent performance metrics over multi-label outputs. The effect of transfer learning was investigated in the experiments by pretraining the model on one dataset and then fine-tuning it on the other. Transfer learning between both datasets saves training time while slightly lowering test accuracy, according to the findings. The data collecting process, also shows promise in the difficulty of transferring models across various settings and configurations. The efficiency of the transfer learning approach is compared across activity recognition datasets such as the MM-Fit Dataset and squat assessment datasets, like the EJUST-SQUAT-21 Dataset and the Single Individual dataset.

In the future, more data can be collected, including more workout exercises, as well as achieving diversity in the subjects participating in the data collection process so that they span a wide range of ages and gender. Such exercises need to be annotated in terms of correctness and the types of mistakes incurred while performing them. Disentanglement (Zhang, Tran, Liu, & Liu, 2020) will be used to extract the relevant features in each performed exercise and make the system more robust to variations in the steady skeletal pose of different subjects in the

Table 3: Results on EJUST-SQUAT-21 Dataset and Single Individual Dataset.

Training Dataset	Pretraining Dataset	Number of Epochs	Accuracy	Type of classification
EJUST-SQUAT-21 dataset	Single Individual Dataset for 60 epochs (Ogata et al., 2019)	83 (end-to-end training)	94% (5-Fold Cross Validation)	multi-label multi-class
	MM-Fit Dataset (Strömbäck et al., 2020) for 25 epochs	38 (fine-tuning)	90.3% (5-Fold Cross Validation)	
		200 (fine-tuning)	88% (5-Fold Cross Validation)	
Single Individual Dataset	-	60 (end-to-end training)	87% (80%-10%-10% training-validation-test split)	multi-class
	-	60 (end-to-end training)	80% (60%-10%-30% training-validation-test split)	
	EJUST-SQUAT-21 for 83 epochs	54 (fine-tuning)	80.2% (80%-10%-10% training-validation-test split)	
	MM-Fit Dataset (Strömbäck et al., 2020) for 25 epochs	120 (fine-tuning)	73% (80%-10%-10% training-validation-test split)	

dataset. A wide range of tasks will need to be done, including the recognition of the exercise, the level of performance, the mistakes done, etc. Transfer learning will play a major role as well in the effectiveness of developing such a virtual coaching system.

ACKNOWLEDGEMENTS

This work is funded by the Information Technology Industry Development Agency (ITIDA), Information Technology Academia Collaboration (ITAC) Program, Egypt – Grant Number (ARP2020.R29.2 - VCOACH: Virtual Coaching for Indoors and Outdoors Sporting).

REFERENCES

- Abou Elmagd, M. (2016). Benefits, need and importance of daily exercise. *Int. J. Phys. Educ. Sports Health*, 3(5), 22–27.
- Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., & Grundmann, M. (2020). Blazepose: On-device real-time body pose tracking. *arXiv preprint arXiv:2006.10204*.
- Bernardino, A., Vismara, C., i Badia, S. B., Gouveia, É., Baptista, F., Carnide, F., ... Gamboa, H. (2016). A dataset for the automatic assessment of functional senior fitness tests using kinect and physiological sensors. In *2016 1st international conference on technology and innovation in sports, health and wellbeing (tishw)* (pp. 1–6).
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., & Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1), 172–186.
- Capecchi, M., Ceravolo, M. G., Ferracuti, F., Iarlori, S., Monteriù, A., Romeo, L., & Verdini, F. (2019). The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(7), 1436–1448.
- Chéron, G., Laptev, I., & Schmid, C. (2015). P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 3218–3226).
- Fayez, A., Sharshar, A., Hesham, A., Eldifrawi, I., & Gomma, W. (2022). Vals: A leading visual and inertial dataset of squats. In *2022 16th international conference on ubiquitous information management and communication (imcom)* (pp. 1–8).
- Füzéki, E., Groneberg, D. A., & Banzer, W. (2020). Physical activity during covid-19 induced lockdown: recommendations. *Journal of Occupational Medicine and Toxicology*, 15(1), 1–5.
- Garber, C., Blissmer, B., Deschenes, M., Franklin, B., Lamonte, M., Lee, I.-M., ... Swain, D. (2011, 07). Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: Guidance for prescribing exercise. *Medicine and science in sports and exercise*, 43, 1334–59. doi: 10.1249/MSS.0b013e318213febf
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Hung, J.-S., Liu, P.-L., & Chang, C.-C. (2020). A deep learning-based approach for human posture classification. *Proceedings of the 2020 2nd International Conference on Management Science and Industrial Engineering*.
- Iqbal, U., Garbade, M., & Gall, J. (2017). Pose for action - action for pose. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 438–445.
- Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018a). End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7122–7131).
- Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018b). End-to-end recovery of human shape and pose. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7122–7131.
- Kaur, H., Singh, T., Arya, Y. K., & Mittal, S. (2020). Physical fitness and exercise during the covid-19 pandemic: a qualitative enquiry. *Frontiers in psychology*, 11, 2943.
- Ke, Q., Bennamoun, M., An, S., Sohel, F., & Boussaid, F. (2017). A new representation of skeleton sequences for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3288–3297).
- Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: a large video database for human motion recognition. In *2011 international conference on computer vision* (pp. 2556–2563).
- Liao, Y., Vakanski, A., & Xian, M. (2020, Feb.). A deep learning framework for assessing physical rehabilitation exercises. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(2), 468–477.
- Lorenzetti, S., Ostermann, M., Zeidler, F., Zimmer, P.,

- Jentsch, L., List, R., ... Schellenberg, F. (2018). How to squat? effects of various stance widths, foot placement angles and level of experience on knee, hip and trunk motion and loading. *BMC Sports Science, Medicine and Rehabilitation*, 10(1), 1–11.
- Martinez, J., Hossain, R., Romero, J., & Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2640–2649).
- Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2017). Monocular 3d human pose estimation in the wild using improved CNN supervision. In *2017 International Conference on 3d Vision (3dV)* (pp. 506–516).
- Nie, B. X., Xiong, C., & Zhu, S.-C. (2015). Joint action recognition and pose estimation from video. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 1293-1301). doi: 10.1109/CVPR.2015.7298734
- Ogata, R., Simo-Serra, E., Iizuka, S., & Ishikawa, H. (2019). Temporal distance matrices for squat classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 0–0).
- Ohuruogu, B. (2016). The contributions of physical activity and fitness to optimal health and wellness. *Journal of Education and Practice*, 7, 123-128.
- Popa, A.-I., Zanfir, M., & Sminchisescu, C. (2017). Deep multitask architecture for integrated 2d and 3d human sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6289–6298).
- Reiner, M., Niermann, C., Jekauc, D., & Woll, A. (2013). Long-term health benefits of physical activity—a systematic review of longitudinal studies. *BMC public health*, 13(1), 1–9.
- Schuch, F. B., Vancampfort, D., Richards, J., Rosenbaum, S., Ward, P. B., & Stubbs, B. (2016). Exercise as a treatment for depression: a meta-analysis adjusting for publication bias. *Journal of psychiatric research*, 77, 42–51.
- Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.
- Strömbäck, D., Huang, S., & Radu, V. (2020). Mmfit: Multimodal deep learning for automatic exercise logging across sensing devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4), 1–22.
- Vybornyi, A. I., Rozaliev, V. L., & Orlova, Y. A. (2017). Controlling the correctness of physical exercises performance. In *Iv international research conference "information technologies in science, management, social sphere and medicine" (itsmssm 2017)* (pp. 233–237).
- Zhang, Z., Tran, L., Liu, F., & Liu, X. (2020). On learning disentangled representations for gait recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.